Counting things & analysing text

Instructions for completing the exercise

You will find the detailed instructions on how to complete this exercise in the Jupyter notebook file counting.ipynb . To start with, you need to download the synthetic data set dataset.jsonl and upload it to the directory that contains the notebook file, so that the code block in the notebook can read the data properly. To reduce the runtime, the notebook file suggests performing the analysis on a random sample of 500 posts for each question (see notebook for details). However, if you like, you are welcome to run the analysis on the full dataset (expected runtime is a few hours).

All the points in this exercise will come from answering the questions on A+, but for some of them you need to extract information from the data set by code. Your implementation will not be assessed; the idea is for you to reach the right conclusion no matter how you implement the analysis.

Exercise description

Now that you have investigated the big data characteristics of the SMC dataset, World Climate Organisation (WCO) folks and you are excited to get started with examining the data to gain insights about what people are discussing within the climate related discourse on SMP. WCO folks are interested in measuring the public opinion on the climate debate based on the SMP data. In your discussions with WCO you heard that often climate activists tend to be positive in their discussions, while deniers sometimes post negative content to disrupt the ongoing conversation. You set out to test this hypothesis about the two user groups (activists and deniers) and to examine the public opinion being voiced on SMP.

Points 15 / 15 My submissions 1 / 1 Deadline Friday, 24 March 2023, 19:00 To be submitted alone

The deadline for the assignment has passed (Wednesday, 5 April 2023, 19:00).

Overall opinion

Question 1 15 / 15

You start by investigating the overall opinion being posted on SMP. For assessing the opinions, you decide to identify the sentiment (positive, neutral, negative) for a sample of 500 posts in the dataset. To get an overall pulse of the discourse opinion, you count the posts with each sentiment label. Which of the following statements best matches your observations?

- Most posts are neutral, with negative posts being a close second and very few positive posts.
- Most posts are neutral with few negative or positive posts.
- **Most posts are negative, with neutral posts being a close second and very few positive posts.**
- Neutral, negative and positive posts are nearly equal in count.

Correct!

Submit

Points 20 / 20 My submissions 1 / 1 Deadline Friday, 24 March 2023, 19:00 To be submitted alone

The deadline for the assignment has passed (Wednesday, 5 April 2023, 19:00).

Group size

Question 1 20 / 20

Seeking to better understand how climate change activists and skeptics are posting about the topic, you attempt to identify the activists and skeptics in the dataset by manually curating a seed set of hashtags for each user group: "#climatecrisis", "#climatejustice" for activists; "#climatehoax", "#globalwarminghoax" for skeptics. You classify the users who post activist hashtags but do not post any skeptics hashtags as activists, and conversely as skeptics. Which of the two groups are larger on SMP?

- Activists
- **Skeptics**

Correct!

Submit

Points 15 / 15 My submissions 1 / 1 Deadline Friday, 24 March 2023, 19:00 To be submitted alone

The deadline for the assignment has passed (Wednesday, 5 April 2023, 19:00).

Group activity

Question 1 15 / 15

Which of the two groups are more active (i.e., post more) on SMP?

- Activists
- **Skeptics**

Correct!

Submit

Points 0 / 10 My submissions 1 / 1 Deadline Friday, 24 March 2023, 19:00 To be submitted alone

The deadline for the assignment has passed (Wednesday, 5 April 2023, 19:00).

Activists' sentiment

Question 1 0 / 10

What is the dominant sentiment of posts containing the activists' seed set of hashtags, namely "#climatecrisis" and "#climatejustice"?

- **Positive**
- Negative
- Neutral

Incorrect

Submit

Points 10 / 10 My submissions 1 / 1 Deadline Friday, 24 March 2023, 19:00 To be submitted alone

The deadline for the assignment has passed (Wednesday, 5 April 2023, 19:00).

Skeptics' sentiment

Question 1 10 / 10

What is the dominant sentiment of posts containing the skeptics' seed set of hashtags, namely "#climatehoax" and "#globalwarminghoax"?

- Positive
- **Negative**
- Neutral

Correct!

Submit

Points 10 / 10 My submissions 1 / 1 Deadline Friday, 24 March 2023, 19:00 To be submitted alone

The deadline for the assignment has passed (Wednesday, 5 April 2023, 19:00).

Sentiment observations

Question 1 10 / 10

Do the above two observations imply that the starting hypothesis ("activists mostly post positive content and skeptics post negative content") holds?

- Yes
- **No**

Correct!

Submit

Points 20 / 20 My submissions 1 / 1 Deadline Friday, 24 March 2023, 19:00 To be submitted alone

The deadline for the assignment has passed (Wednesday, 5 April 2023, 19:00).

Hypothesis

Question 1 20 / 20

If you identify the dominant sentiment for ALL the posts made by activists and skeptics (not just the ones containing the seed set of hashtags) using a sample of 500 posts, does the starting hypothesis hold?

- Yes, activists mostly post positive and skeptics mostly post negative posts
- **No, activists mostly post neutral and skeptics mostly post negative posts**
- No, both activists and skeptics mostly post neutral posts
- No, both activists and skeptics mostly post negative posts

Correct!

Submit